Computerunterstützte Textanalyse

Messelken, Hans

Historical Social Research, Vol. 14 — 1989 — No. 4, 86-93

Computerunterstützte Textanalyse

*Hans Messelken**

Voraussetzungen

Ziel der Analysen sind objektivierte und differenzierte Beschreibungen von Texten und ihrer Verständlichkeit. Diese ergibt sich aus Länge und Differenzierungsgrad der Textoberfläche. Sie wird als Maßzahl VI ausgedrückt und aus Anzahl, Länge sowie Beziehungsdichte von Buchstaben, Wörtern und Sätzen berechnet.

Anzahl und Länge von sprachlichen Zeichen sind meßbare Größen. Nicht meßbar, aber objektiviert zu beschreiben ist die Beziehungsdichte des Textes, also sein Reichtum an sprachlichen Verknüpfungen.

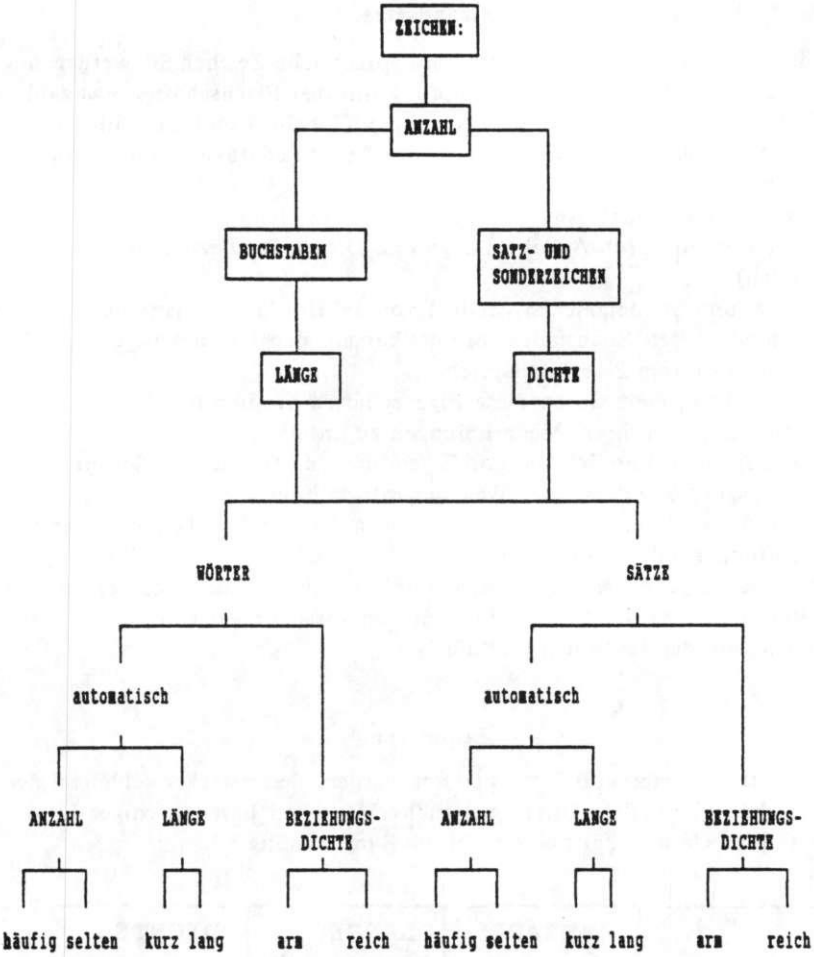
Die Textstatistik (Kapitel 2) erfaßt die meßbaren Teilbereiche, die Strukturanalyse, die objektivierten Beschreibungen (Kapitel 3) und die Dokumentation vollständiger Listen und Belegsammlungen (Kapitel 4). In allen Feldern geht es um

- allgemein anerkannte Regeln aus Grammatik und Stilistik (Duden-Norm) sowie
- Vergleiche mit anderen Texten. Diese sind um so wichtiger, je weniger plausible Normen für die Beschreibung von Texten zur Verfügung stehen.

Zweck der Textanalyse ist es, Normen an der sprachlichen Oberfläche zu beschreiben und Abweichungen auf zugrundeliegende kommunikative Absichten zu untersuchen. Da es bislang keine hinreichenden Normen gibt, ist der Einzelwert in ein Profil vergleichbarer Größen innerhalb des Textes zu differenzieren; dabei geht es um die stilistische Bedeutung, wie sie in Dichte, Steilheit und Textdeckung erfaßt wird (Punkt 2.01). Daraus ergeben sich Vergleichsmöglichkeiten mit anderen Texten (Punkt 2.02). Kommunikative Absichten äußern sich in stilistischen Signalen, soweit diese an der Oberfläche von Texten objektiviert nachweisbar sind (Punkt 3). Nur der gesammte Zusammenhang ermöglicht Aussagen zum semantischen oder stilistischen Zweck der untersuchten Merkmale.

»Verständlichkeit« gilt in der Textstatisktik also als Hinweis auf die Differenzierung in der jeweiligen Zeichenklasse (Graphik A).

* Address all Communications to Hans Messelken, Seminar für Deutsche Sprache und ihre Didaktik, Universität zu Köln, Erziehungswissenschaftliche Fakultät, Albertus Magnus-Platz, 5000 Köln 41, FRG



Graphik A: Zeichen und Eigenschaften

- Unterschiedliche sprachliche Zeichen sind schwieriger als wiederholte.
- Viele Zeichen sind schwieriger als wenige.
- Lange Zeichenfolgen sind schwieriger als kurze.
- Dichte, beziehungsreiche Zeichenfolgen sind schwieriger als beziehungsarme.
- Satzformen sind schwieriger als Wortformen.

Historical Social Research, Vol. 14 — 1989 — No. 4, 86-93

Grundsätze

Buchstaben, Wörter und Sätze sind sprachliche Zeichen. Sie werden in der Textstatistik durch die Verknüpfung mit den Eigenschaften »Anzahl, Länge und Beziehungsdichte« beschrieben (Tabelle 1 und Graphik A):

- Buchstaben werden allein durch ihre Anzahl charakterisiert (wenige - viele),
- Wortformen durch Anzahl und Länge (kurz • lang),
- Satzformen durch Anzahl, Länge und Dichte der Beziehungen (arm - reich),
- Satz- und Sonderzeichen (Zeile 1 von Tabelle 1) verweisen im Unterschied zu den Buchstaben unmittelbar auf den Differenzierungsgrad, wie dies ihrem Zweck entspricht.

Diese konkreten, durch diese Eigenschaften bestimmten Zeichen sind systematisch von ihren Wiederholungen zu unterscheiden:

- Unterschiedliche Zeichen und Eigenschaften spiegeln den Differenzierungsgrad und damit den Wert der Information.
- Die Wiederholung von Zeichen führt in der Regel zu keiner Differenzierung, sondern zur Redundanz.

Daraus ergeben sich Anhaltspunkte zur Abschätzung des Schwierigkeitsgrades, also der Verständlichkeit von sprachlichen Oberflächen, soweit sie mit der Textstatistik erfaßt ist:

Erläuterung

Je öfter Wörter und Sätze benutzt werden, desto stärker schleifen sie sich ab: Sie enthalten für immer mehr Menschen immer weniger Informationen. Daraus läßt sich eine grobe Rangordnung bilden:

	ANZAHL	LÄNGE	DICHTE
leicht	häufig	kurz	allgem.
schwierig	selten	lang	speziell

Die leichtesten Sprachformen sind häufig und kurz; sie sind ziemlich allgemein. Die schwierigsten sind im Vergleich dazu selten und lang; sie sind recht speziell. Ob diese Sprachformen Wörter oder Sätze sind, spielt zunächst keine Rolle. Vielmehr ist dies anscheinend eine Frage der individuellen Neigung: Nur sehr selten benutzt jemand sehr lange Wörter und gleichzeitig höhst komplexe Satzformen: Der eine differenziert lieber le-

xikalisch, der andere syntaktisch. Für sehr feine Differenzierungen benötigt man Wortschatz und Satzbau; freilich wird der Text in diesem Fall ausgesprochen schwierig.

Eine Sonderrolle spielt die Wiederholung: Sie erhöht zwar die Anzahl, nicht aber den Differenzierungsgrad sprachlicher Formen.

Wenige kurze und beziehungsarme Wörter in wenigen und beziehungsarmen Sätzen sind leichter verständlich als viele lange und beziehungsreiche Wörter in entsprechenden Sätzen: Die zehn Gebote sind leichter verständlich als Bürgerliches- und Strafgesetzbuch. Freilich sind hier kaum vergleichbare Aufgaben in ganz unterschiedlichen Umfeldern zu erfüllen: Der Wunsch nach Verständlichkeit und die erreichbaren Möglichkeiten des einzelnen Textes sind nicht immer auf einen Nenner zu bringen. Der notwendige Kompromiß ergibt sich aus der Strukturanalyse.

Ebenen der Textanalyse

STA berücksichtigt die drei in Tabelle 1 skizzierten Ebenen. Die Zeilen betreffen die sprachliche Oberfläche: Zeichen, Buchstaben, Wörter, Sätze und den Text - soweit sich dies alles überhaupt mit Maßzahlen vergleichen und in Listen erfassen läßt. In den Spalten erscheinen die zugehörigen Eigenschaften (Anzahl, bzw. Häufigkeit, Länge und Dichte). Sie beziehen sich auf Maßzahlen und Listen.

In der ***Texisunistik*** werden Maßzahlen zu statistisch erfaßbaren Größen zusammengestellt, kommentiert, auf den Textbefund bezogen und bewertet. Hier ist der Informationsrahmen zwar recht eng und abstrakt, dafür aber ziemlich zuverlässig. Graphik A zeigt, welche Merkmale automatisch berechnet werden.

Die *Strukturanalyse* betrifft den Zweck des Textes, die Absicht des Verfassers und die dabei verwendeten Stilmittel. Was der Absender der sprachlichen Botschaft wirklich im Sinn hat, läßt sich nur indirekt ermitteln: Der Text enthält an seiner Oberfläche Signale, die in Funktions-, Intentions- und Stilanalyse systematisch und objektiviert zu vergleichen sind.

Erfaßt werden Gestaltung der Seite, Anzahl und Beziehungsdichte der Zeichen, Stichwörter zum Inhalt sowie sprachliche Funktionen, Absichten des Verfassers und stilistische Muster. Damit wird eine Fülle von Informationen gewonnen, die abstrakt genug sind, um einen Überblick zu bieten und hinreichend konkret, um für Beurteilung und Bewertung eine anschauliche und objektivierte Grundlage zu stellen. Die hier erforderlichen Zuordnungen sind und bleiben subjektiv; verschiedene Bearbeiter werden zu mehr oder weniger abweichenden Ergebnissen kommen. Dieser unvermeidliche Effekt wird durch eine Vielzahl von Einzelentscheidungen, Quer- und Kontrollbezügen gemildert.

Die *Dokumentation* liefert vollständige Listen zu bestimmten Fragen, die sich weniger gut für den statistischen Zugriff eignen. Damit gewinnt man zwar mehr und konkretere, allerdings auch weniger übersichtliche Informationen.

Quantitative Eigenschaften von sprachlichen Zeichen

Die in der Textstatistik erfaßten Eigenschaften von sprachlichen Zeichen sind nach den o.a. Grundsätzen rein quantitativ. Sie bleiben an der sinnlich wahrnehmbaren Textoberfläche, die automatisch nach diesen Eigenschaften abgefragt wird. Das geschieht mit dem Betriebssystem MS DOS (Version 3.2) auf marktüblichen Bürorechnern mit ausgebauten Speichern und Festplatten.

- Die Grundlage der automatischen Analyse bildet der »Worderuncher - Text Indexing and Retrieval Software« in der Version 4.21, Electronic Text Corporation, Brigham Young University, Provo, Utah.
- Zusatzroutinen zu Wort- und Satzstatistik wurden von G. Frackenpohl entwickelt.
- Darüber hinaus werden übliche Programme der Bürokommunikation verwendet:
- Wordstar 3.4 und WordPerfect 4.2 zur Textverarbeitung,
- Primus 2.0, Softex, Saarbrücken 1987, zur Prüfung der Rechtschreibung und Ermittlung der unteren Grenze des speziellen Wortschatzes,
- Multiplan 3.0, Microsoft 1987, zur Berechnung von Teilauswertungen,
- Chart 2.0, Microsoft 1987, zur graphischen Darstellung einzelner Befunde.

Entsprechend fallen die Befunde aus. Sie sind zwar objektiv, vollständig und differenziert, aber sehr inhaltsarm. Dafür werden die eben aufgestellten Grundsätze (Punkt 1.2) konsequent eingehalten:

Der Informationswert eines Textes wird aus Anzahl, Länge und Beziehungsdichte der sprachlichen Zeichen ermittelt und mit dem Differenzierungsgrad auf der jeweiligen sprachlichen Ebene gleichgesetzt. Die folgenden Stichwörter zu den jeweiligen Eigenschaften sollen andeuten, aus welchen Gründen die Befunde auf den ersten Blick zwar nichtssagend erscheinen, trotzdem aber keineswegs bloße Binsenweisheiten liefern.

Normen und Abweichungen äußern sich nicht zuletzt in der Häufigkeit der zu analysierenden Zeichen und Merkmale (häufig - selten). Die Anzahl von Buchstaben, Wörtern und Sätzen wird in der Textstatistik, die von Zeichen und Fehlern in der Strukturanalyse ermittelt. Die reine Anzahl reicht freilich nicht einmal zur quantitativen Bestimmung: Auf allen Ebenen ist der Anteil unterschiedlicher Formen ins Verhältnis zu ihren Wiederholungen zu setzen:

- Unterschiedliche Zeichen differenzieren ein Teilsystem (Satz und

Sonderzeichen, Buchstaben, Wörter, Sätze). Sie erhöhen damit den Informationswert - freilich auch den Schwierigkeitsgrad.
- Wiederholungen verfestigen das Teilsystem und erhöhen die Redundanz. Dies senkt den Schwierigkeitsgrad.

Das zentrale Merkmal besteht hier in der Länge der sprachlichen Zeichen (kurz - lang). Lange sprachliche Formen sind differenzierter, damit auch (mehr oder weniger) informationsreicher als kurze. Durch Wechsel zur höheren Ebene werden die Zeichen kürzer, ohne an Differenzierungswert zu verlieren. Darin liegt die Begründung für den Verständlichkeitsindex, der lediglich die Länge von Zeichenfolgen berücksichtigt. Lange Formationen können bekanntlich durch besser verständliche kurze umschrieben werden.

Die Länge von Wörtern und Sätzen wird in der Textstatistik als VI (Wort, Satz, Text) berechnet. Sie kann in der Dokumentation durch Listen überlanger und besonders kurzer Wörter oder Sätze ergänzt werden.

Das zentrale Merkmal liegt im lexikalischen, bzw. syntaktischen Differenzierungsgrad (allgemein = wenig differenziert, speziell = stark differenziert). Die Beziehungsdichte des Textes wird als Maßzahl durch TTR-Wert und Satzdichte-Index ausgedrückt. Für den Wortschatz kommen dazu Listen von »Types«, Schlüssel- und SpezialWörtern sowie ein Volltextindex. In bezug auf Satzbau und Text gibt es Listen überlanger und kurzer Sätze sowie zu Seitengestaltung, Inhalt, Zweck, Absicht und Stil.

Die *Maßzahlen* sind nach Möglichkeit auf 100 normiert, damit man das Gewicht des Befundes besser abschätzen kann. Sie werden in der Textstatistik aufgeführt.

Reicht eine Maßzahl für einen genauen Befund nicht aus, wird sie durch besondere *Listen* in der Dokumentation ergänzt. Die Listen dienen zur genauen Beobachtung von Eigenschaften und Erscheinungsweisen der sprachlichen Formen.

Grenzen

Der Sangesmeister Beckmesser versucht bekanntlich, individuelle Bekundungen über ein- und denselben Leisten zu schlagen, indem er wenige normative Merkmale mechanisch anwendet und danach Lob und Tadel handwerklich gerecht, gleichsam automatisch zumißt. In gewisser Weise geschieht bei STA etwas Ähnliches. Die Grundregeln lauten deshalb:
- Abweichungen von normativen Merkmalen wie etwa dem VI TEXT werden zunächst einmal ermittelt.
- Die Verständlichkeit läßt sich nach STA nicht durch einen einzelnen Wert, sondern nur in einem Profil unterschiedlicher Merkmale erfassen.
- Das einzelne Merkmal gewinnt erst im Vergleich mit allen anderen

TEXTANALYSE : ÜBERBLICK

	ANZAHL	LÄNGE	BEZIEHUNGSDICHTE
ZEICHEN	Gesamt: 2.4.2		Satzdichte: 3.3.1
BUCHSTABEN	Gesamt: 2.01		
WÖRTER Maßzahlen: Listen:	2.3 und 4.2 Basisw.: 2.3.6 Verschied.2.3.2 Wiederhol.2.3.4	4.4 VI-WORT: 2.2.2 Überlange: 2.3.5 Kurzwörter:2.3.5 Abkürzungen: 4.4;	4.3 TTR: 2.3.3 Verschiedene: 2.3.2 Schlüsselwört.: 3.5 Volltextindex: 4.6 Spezialwörter:2.3.7
SÄTZE Maßzahlen: Listen:	Gesamt: 2.4.1	2.4.1 VI-SATZ: 2.2.3 Langsätze, 2.4.1, Kurzsätze: 3.3.1	2.4.2 und 3.3.1 Satzdichte: 2.4.2
TEXT Maßzahl: Listen:	3.2 und 4.1 Belegtext: 4.1 Fehler: 3.3.2	VI-TEXT: 2.2.1	4.1 Seitengest.: 3.2 Bez.dichte: 3.3,3.4 Kernbegriffe: 3.4 Inhalt: 3.5 ,4.5 Zweck: 3.6 Absicht: 3.7 Stil: 3.8

Tabelle 1: Maßzahlen und Listen
Die Ziffern beziehen sich auf die Kapitel 2 und 3.

eine gewisse Bedeutung.

- Textstatistik, dreidimensionale Strukturanalyse und Dokumentation sind konsequent zu verbinden.

Erst in der Interpretation kann sich erweisen, welchen semantischen oder stilistischen Zweck die untersuchten Merkmale erfüllen. Das Ziel von STA ist es keineswegs, individuelle sprachliche Bekundungen zu be beck--

messern, sondern Normabweichungen auf ihre kommunikative Zweckmäßigkeit zu untersuchen.

Die Schwierigkeit derartiger Untersuchungen liegt darin, zwischen Umfang und Inhalt der Grundbegriffe das richtige Gleichgewicht zu finden: Nähert man sich der sprachlichen Oberfläche allzusehr, kann man vor lauter Bäumen den Wald nicht mehr sehen (Dokumentation).

Entfernt man sich zu weit, muß man zwischen Binsenweisheit und Vorurteil lavieren, ohne recht beurteilen zu können, was man eigentlich wahrnimmt (Textstatistik). So gibt beispielsweise der Verständlichkeitsindex in der Textstatistik zwar einen allgemeinen Eindruck von der ungefähren »Preislage« des Textes; aber es bedarf der Differenzierung durch zahlreiche andere Größen, damit der Befund überhaupt der Rede wert ist (Strukturanalyse).

Was die Begriffe an Umfang gewinnen, das verlieren sie an inhaltlicher Anschaulichkeit. Der Aufbau von STA soll den erforderlichen Ausgleich zwischen notwendiger Abstraktion und hinreichender Differenzierung erleichtern: Um jederzeit schnellen und klaren Überblick zu bieten, beginnt jede Schrittfolge abstrakt und wird dann zunehmend konkreter.

Das System von Textanalysen kann und soll die Interpretation natürlich nicht ersetzen. Aber es kann sie erleichtern, weil die statistischen und formalisierten Befunde der Interpretation eine vielseitige und zuverlässige Grundlage liefern.